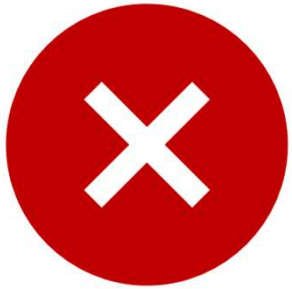
Analisi del trasporto ferroviario italiano

Spolaor Andrea – Luglio 2025

Domande

1. Come sono distribuite le stazioni ferroviarie in Italia?
2. Quali sono le regioni che presentano ritardi medi più alti? E quali ritardi minori?
3. Quali sono state le giornate con più cancellazioni? E quelle con più ritardi?
4. Quali sono le tratte con maggior ritardo medio?

Base dati



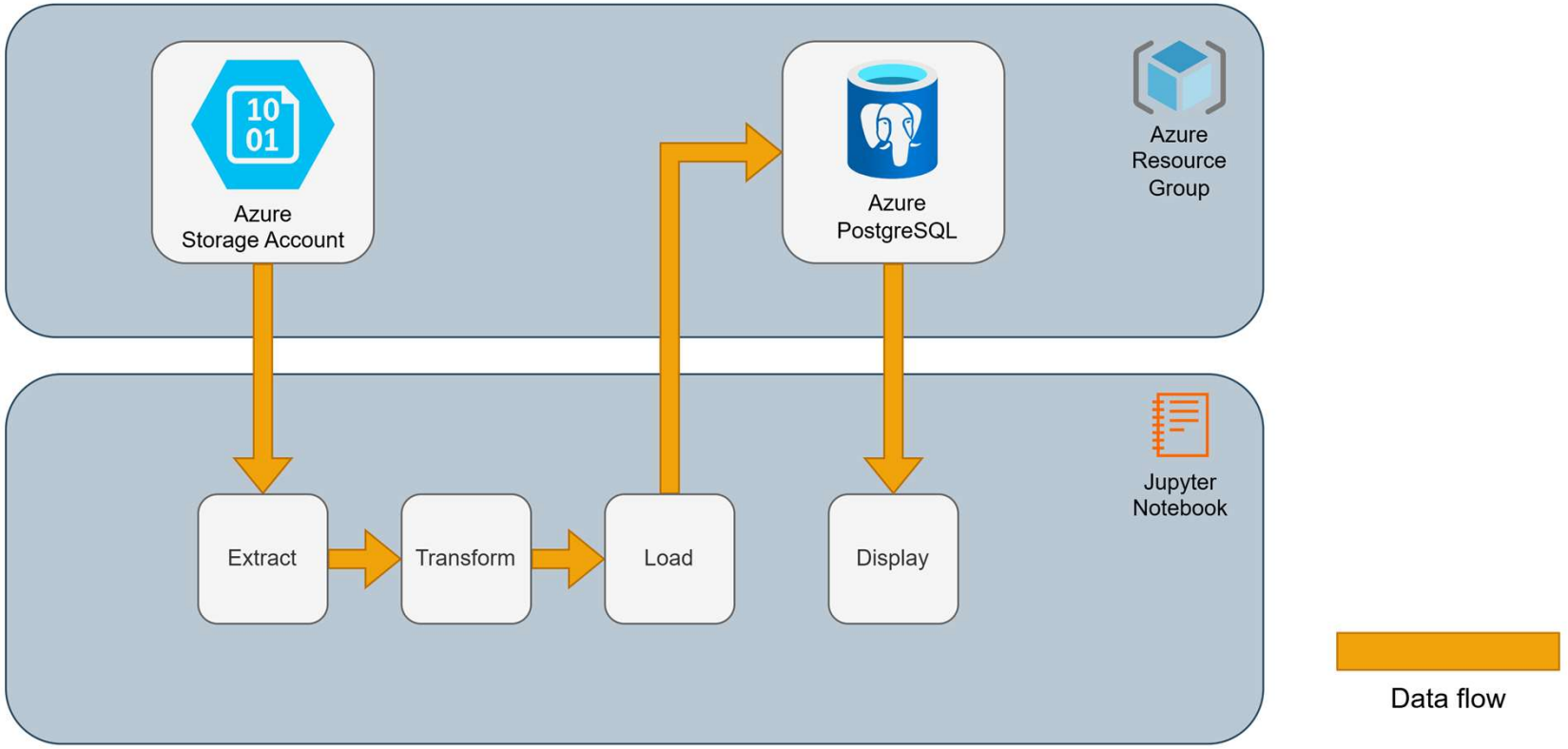
ViaggioTreno: API ufficiali di Trenitalia

- PRO:
  - Dati ufficiali Trenitalia
  - Altissimo numero di informazioni (tempi, meteo, avvisi)
- CONTRO:
  - Intuitività delle API molto bassa
  - Non è possibile interrogare lo storico dati

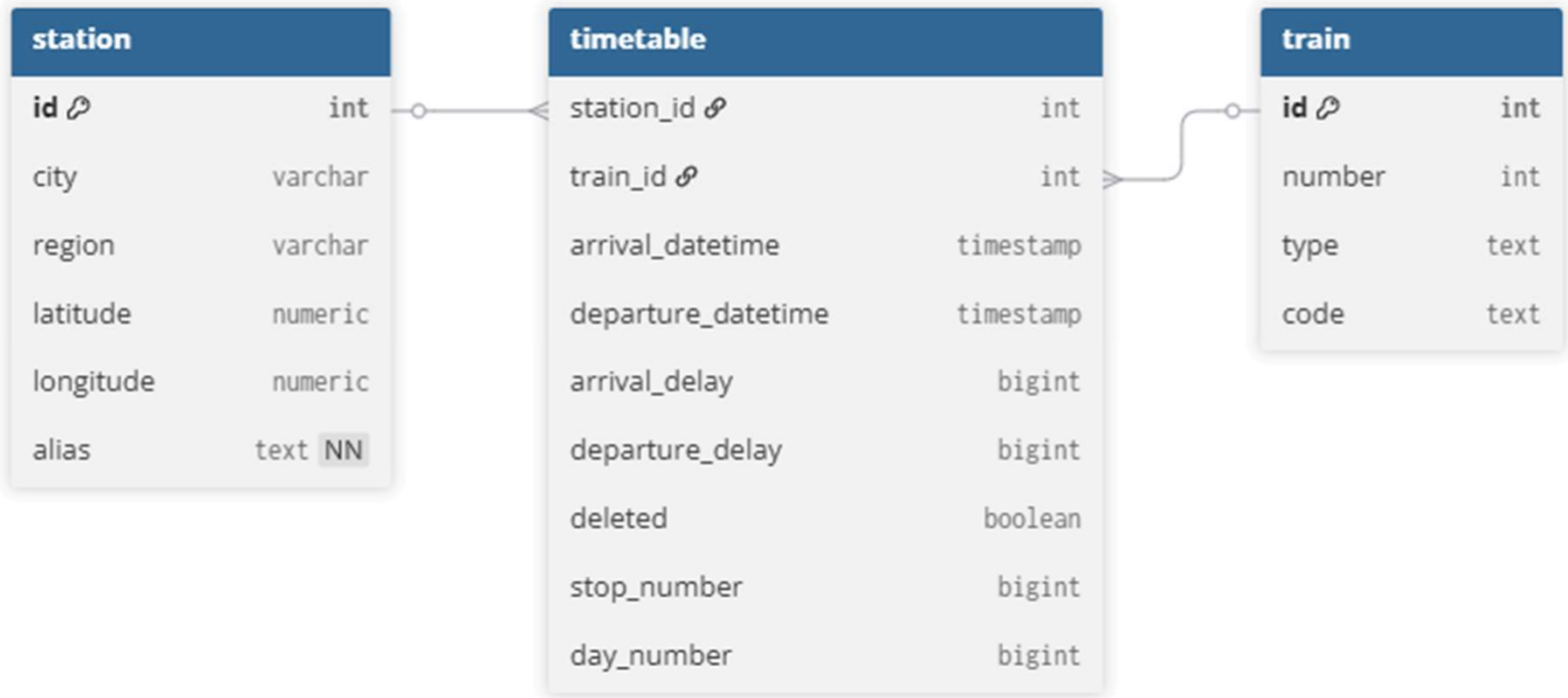
TrainStats: File JSON esposti pubblicamente su DropBox

- PRO:
  - Grande base dati (Luglio 2024 – Giugno 2025)
  - Poca frammentazione dei dati, un file JSON per giorno
- CONTRO:
  - Dati NON normalizzati (es. nome stazioni NON univoche «VENEZIA S.LUCIA», «VENEZIA SANTA LUCIA»)
  - Necessario estrarre informazioni da un JSON complesso

Architettura



Struttura Database



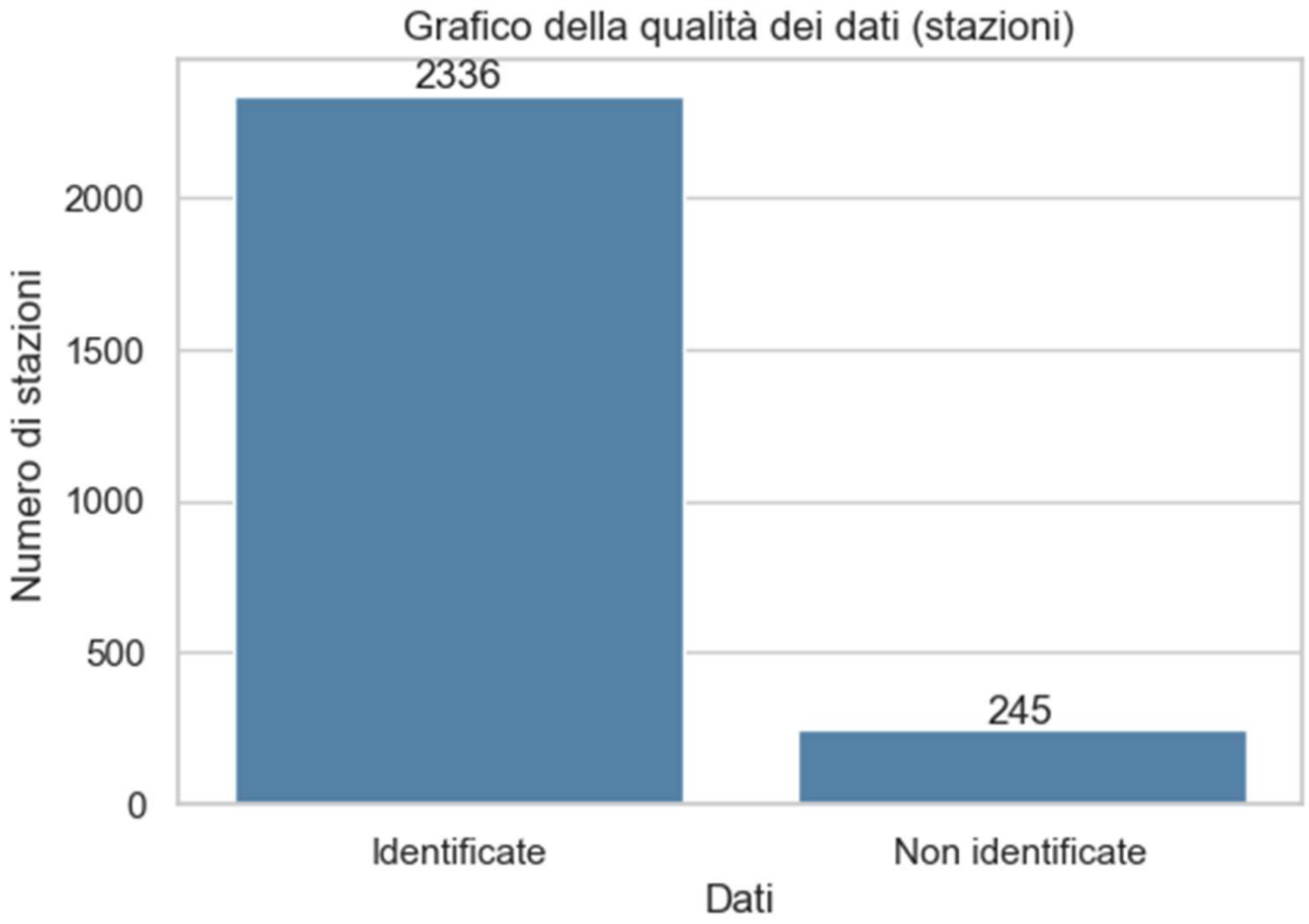
Presentazione dati

Quantità dati in esame

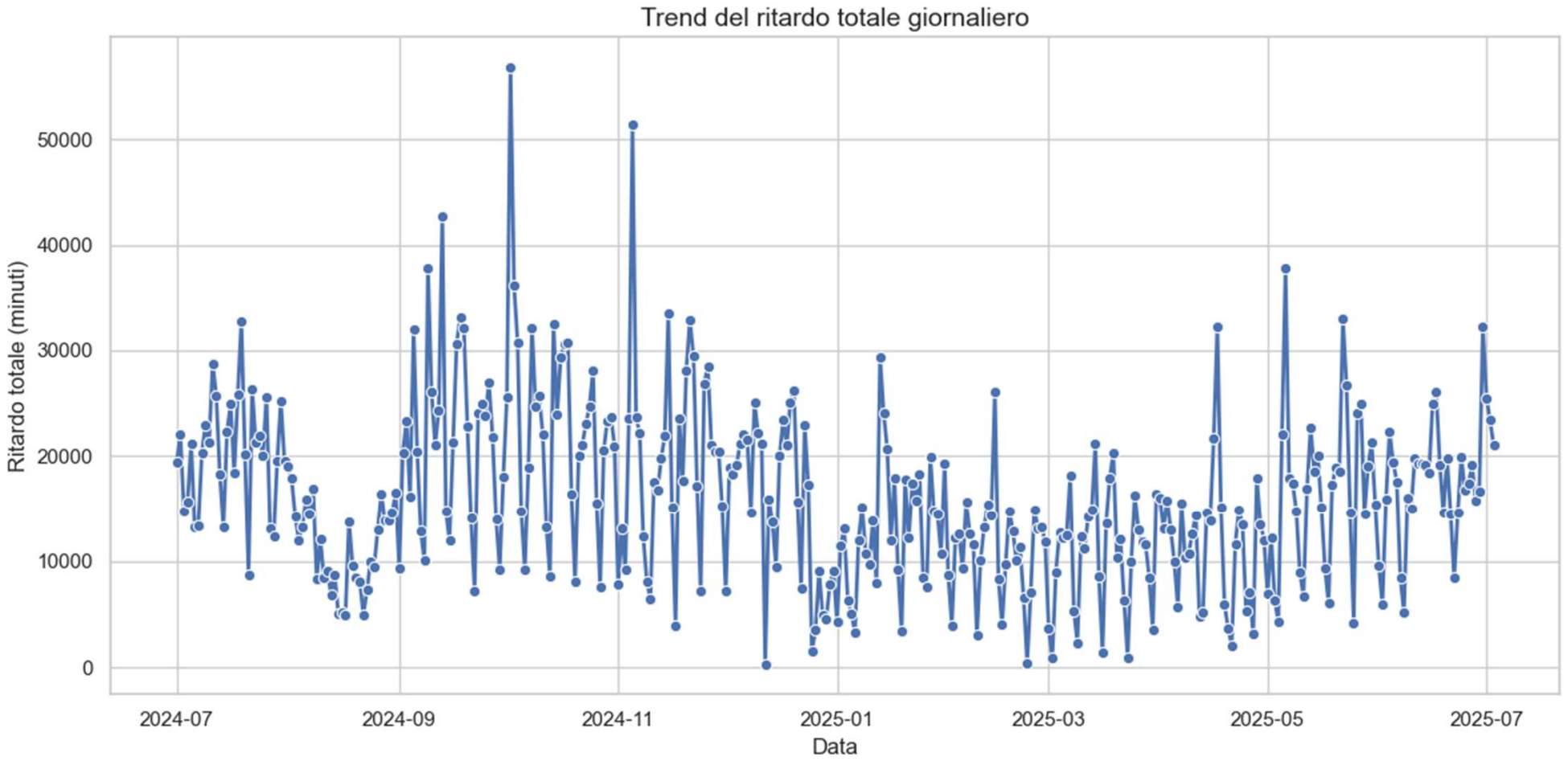
Periodo in esame: Luglio 2024 –Giugno 2025

Entità	Numero righe
Stazioni ferroviarie	2.581
Tratte	18.668
Fermate	33.707.901

Qualità dei dati: Stazioni identificate geograficamente



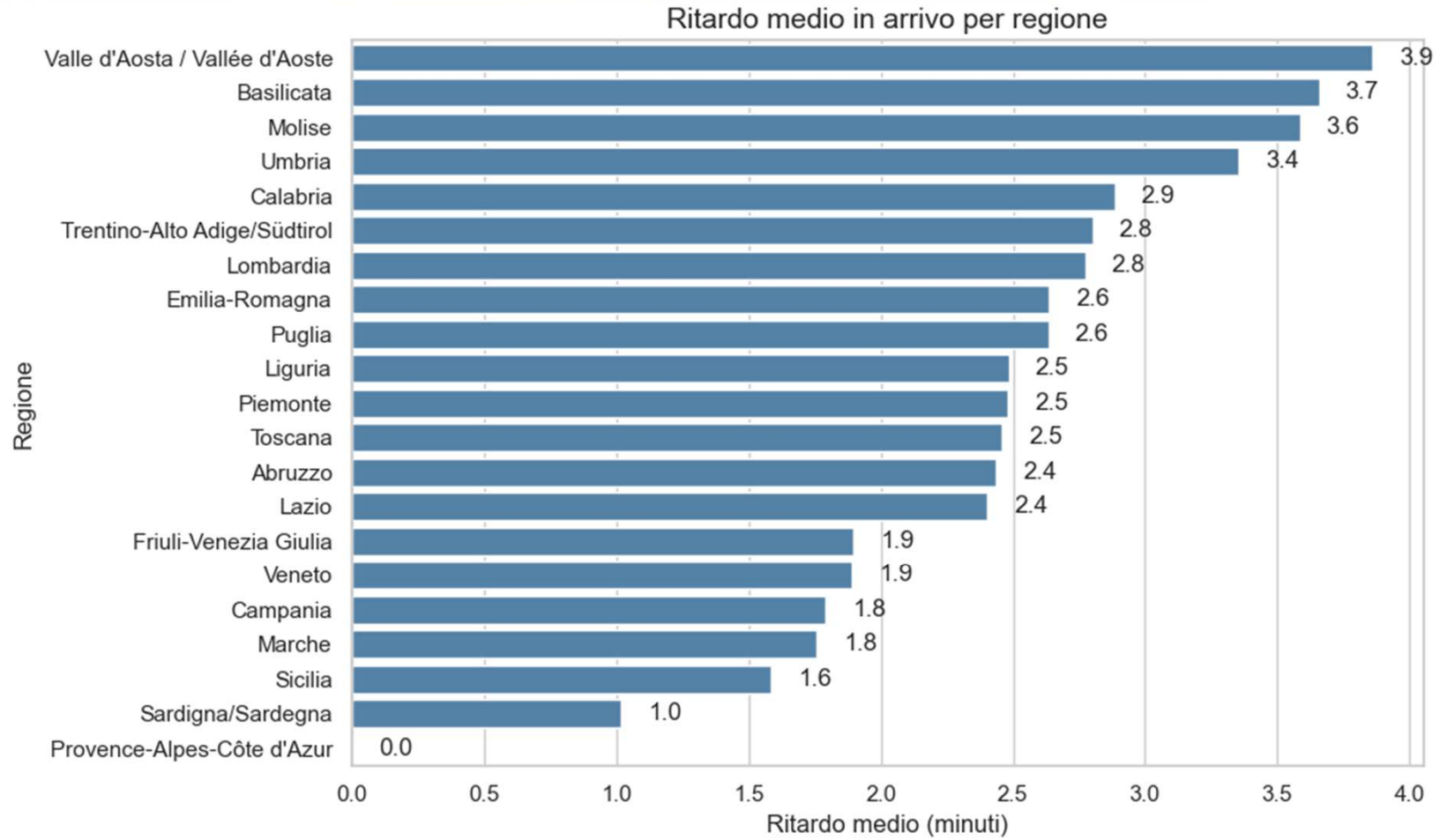
Trend ritardo nel periodo selezionato



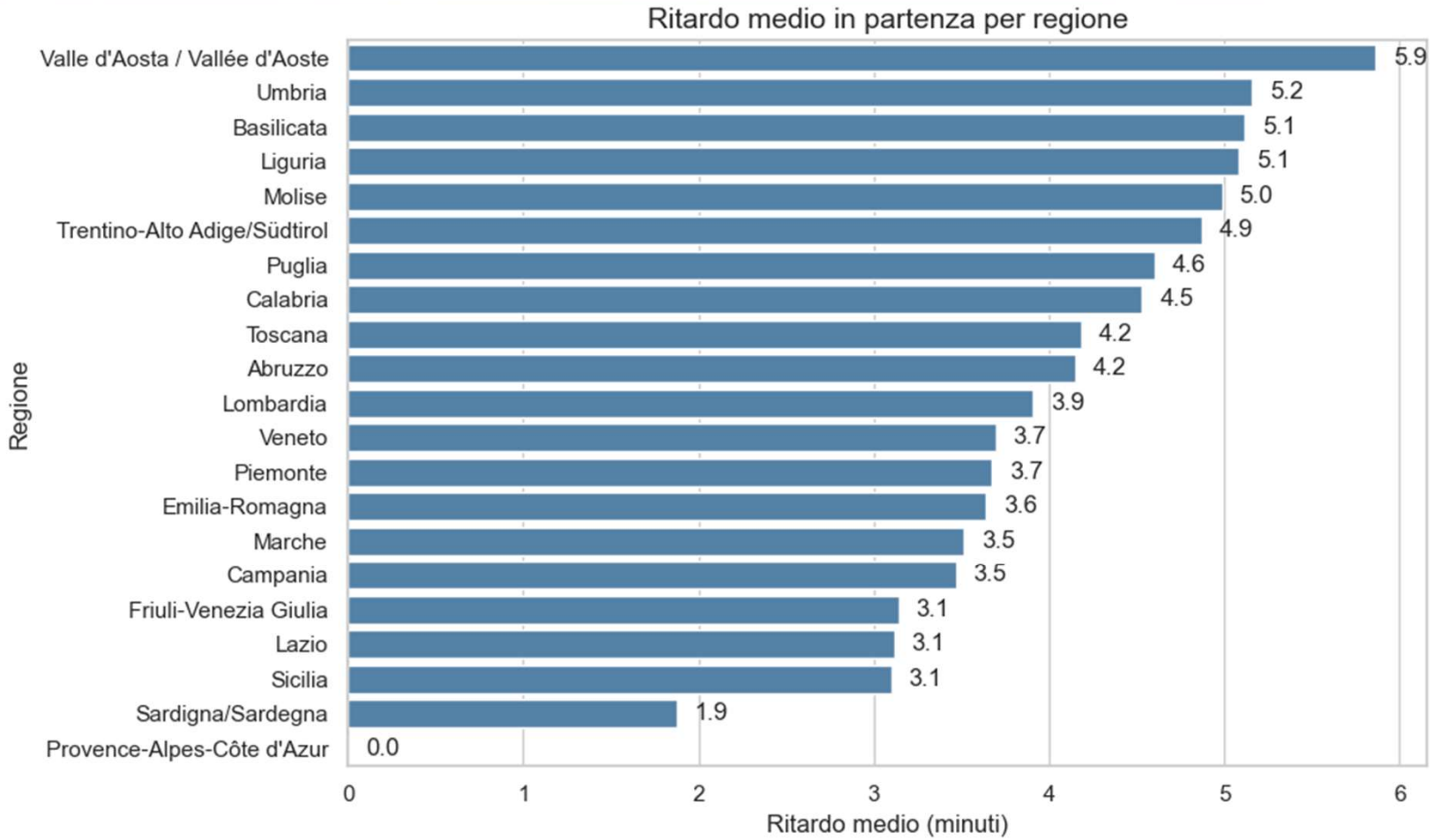
Distribuzione stazioni nel territorio italiano

Stazione ferroviaria

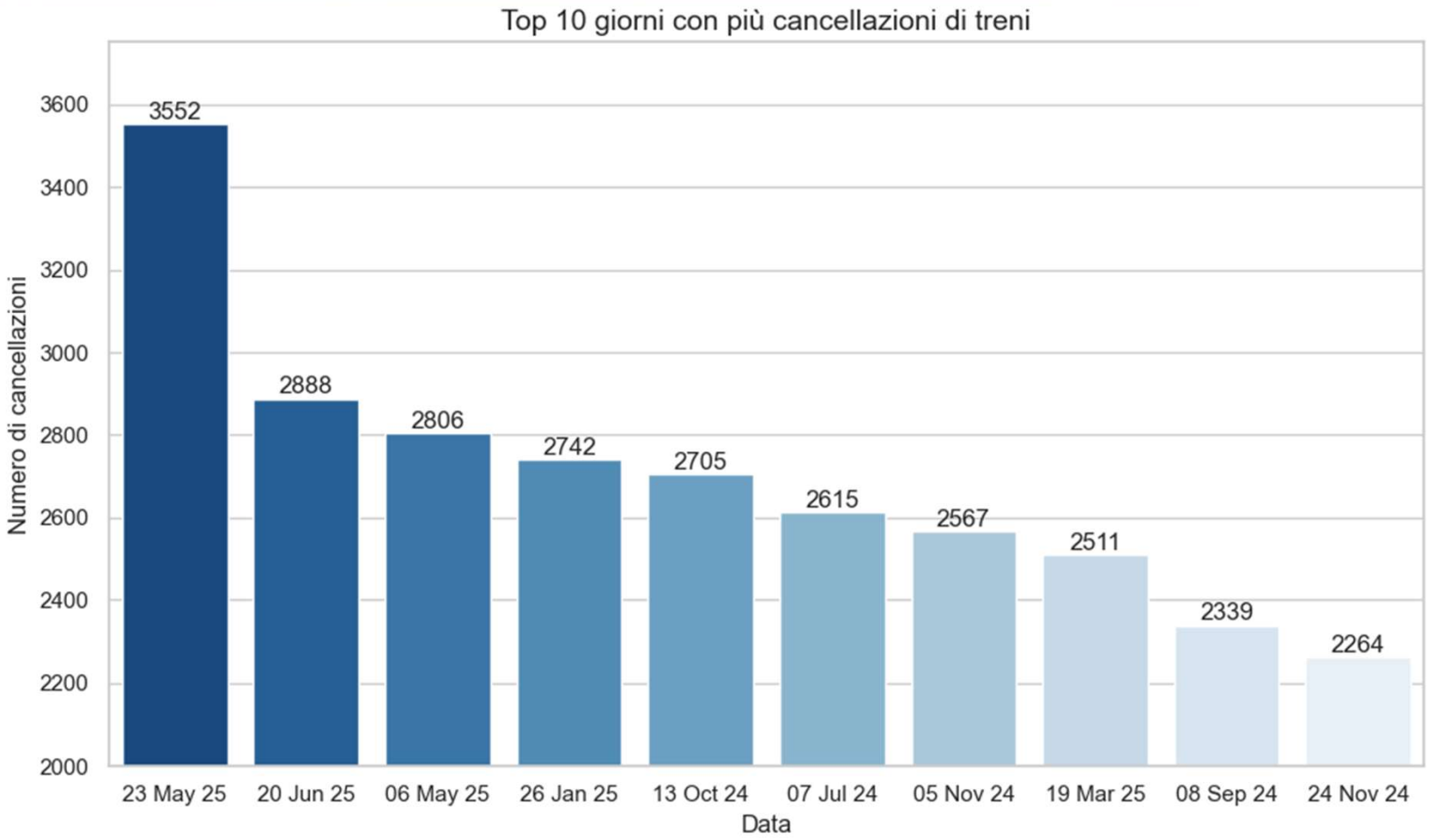
Ritardo medio in arrivo per regione



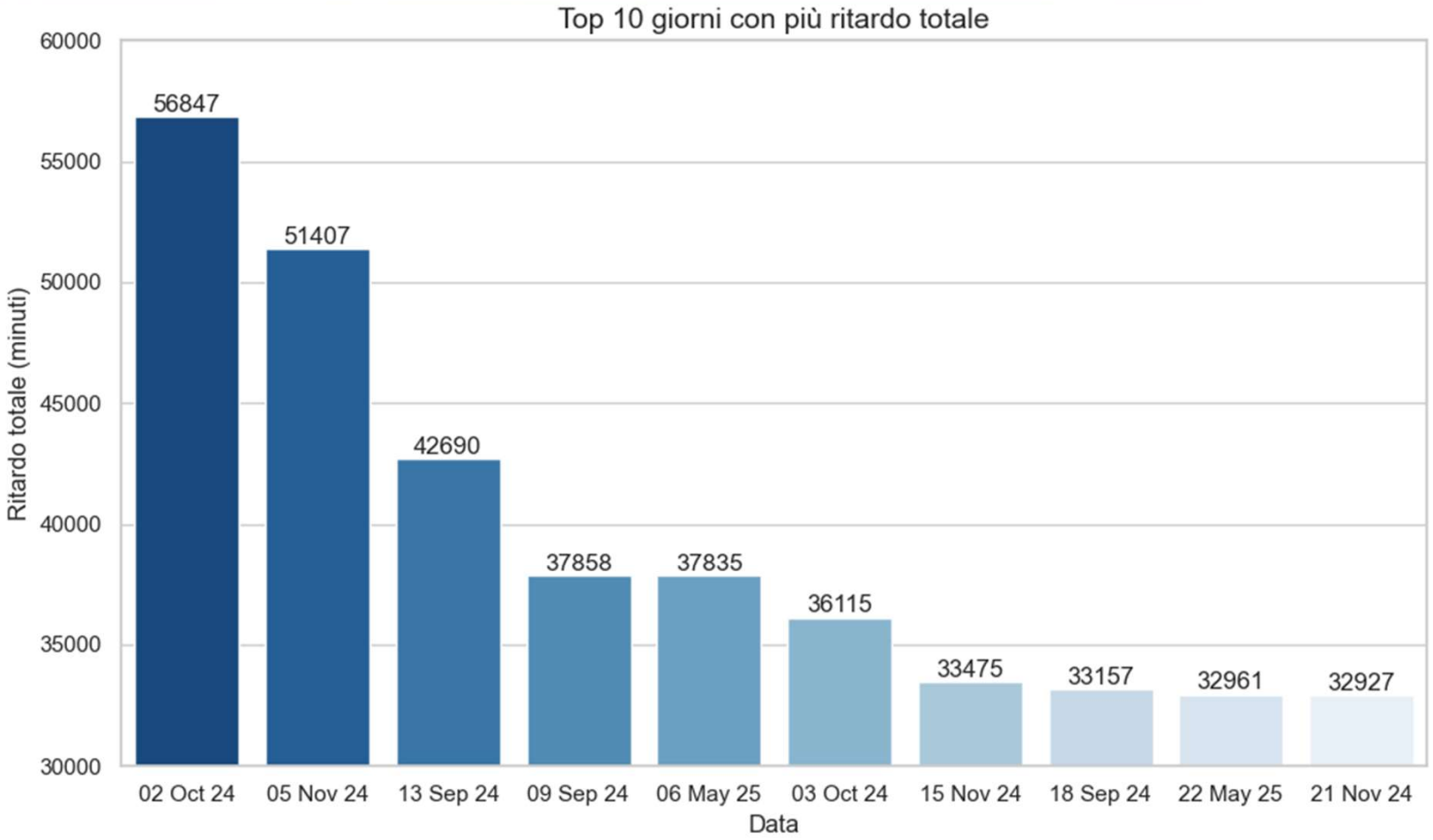
Ritardo medio in partenza per regione



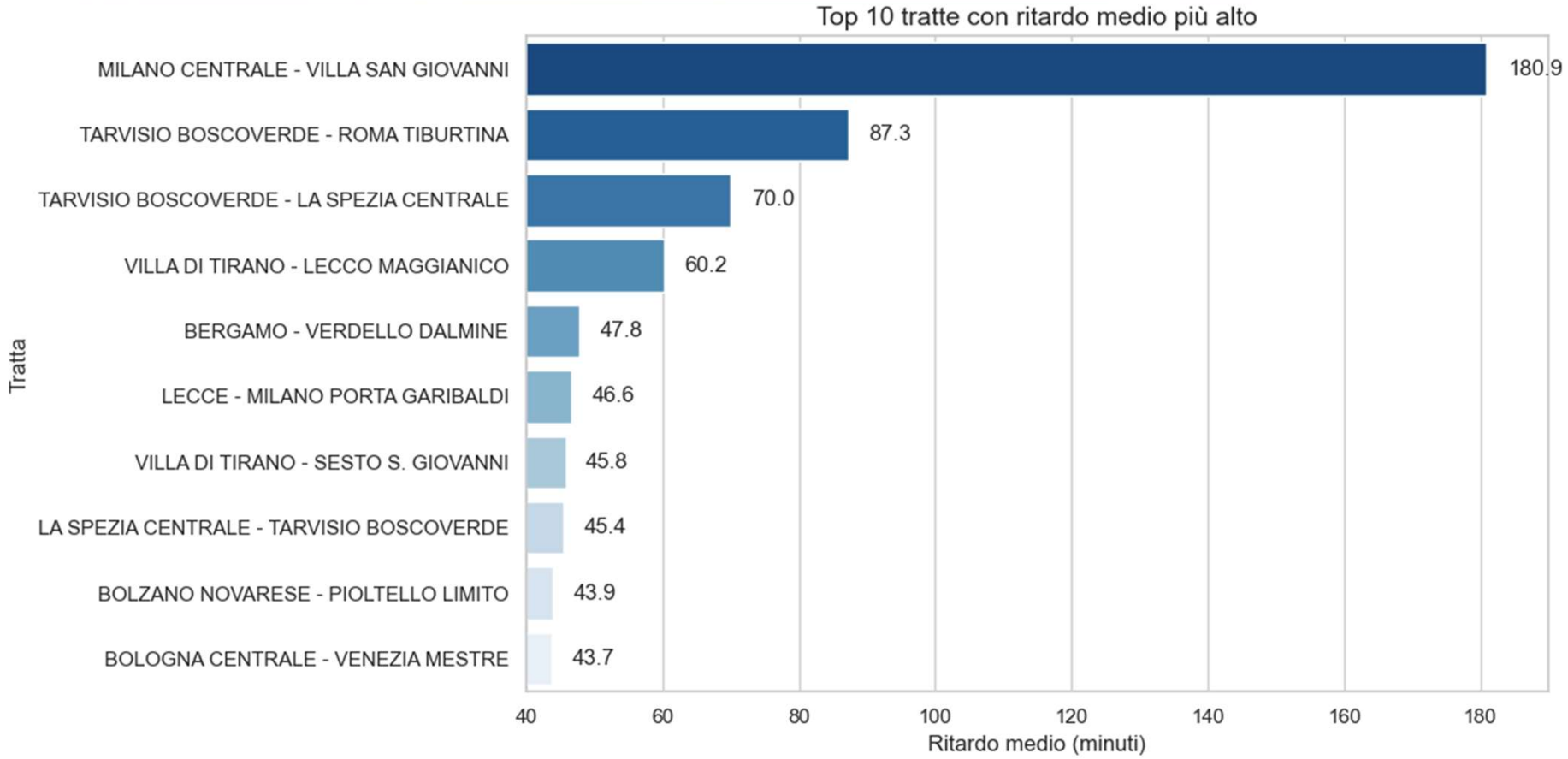
Giorni con maggior numero di cancellazioni



Giorni con maggior minuti di ritardo



Tratte con maggior minuti di ritardo medio*



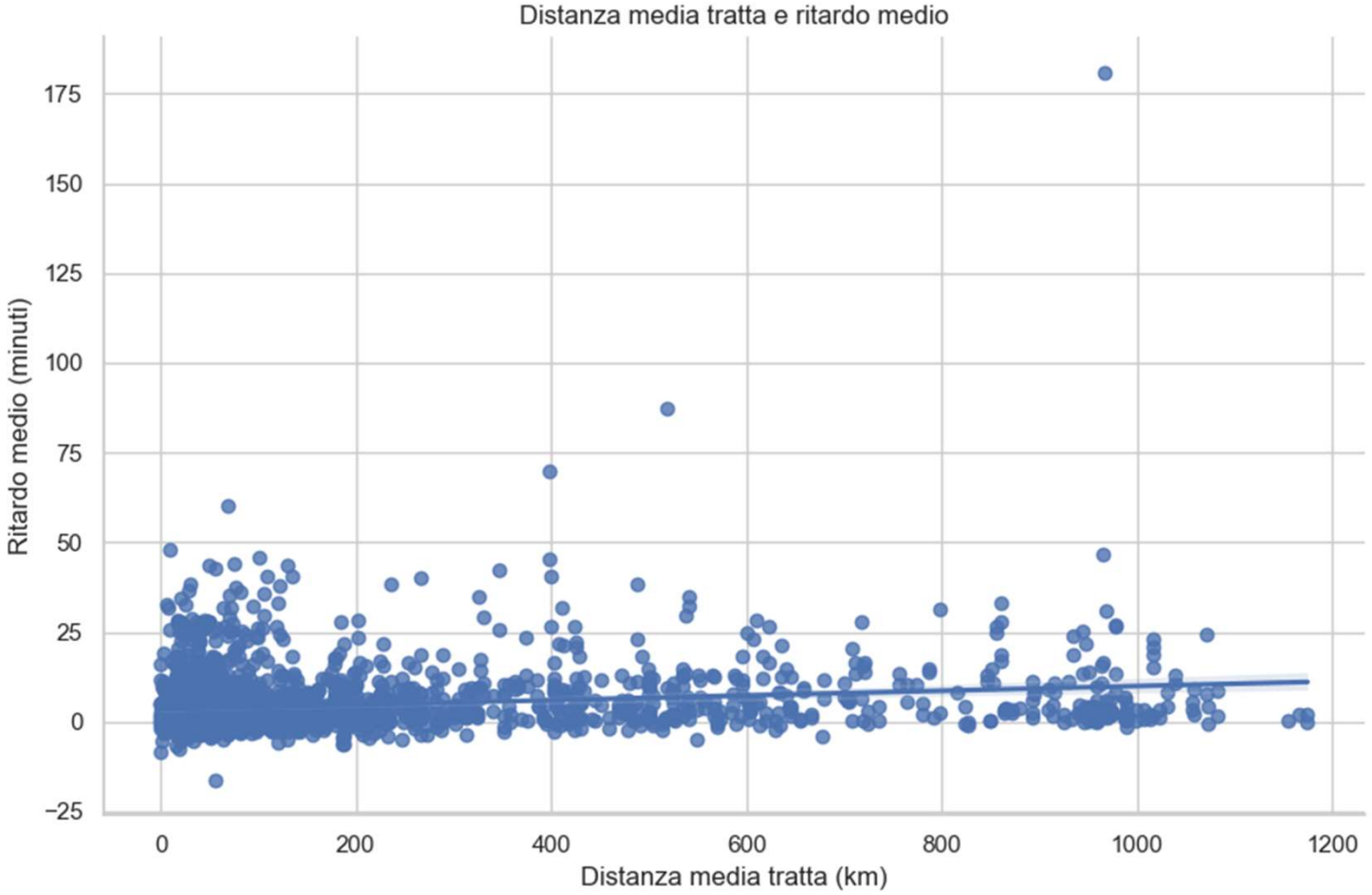
*Tratte percorse almeno 10 volte

Nuove domande:

- C’è correlazione (positiva) tra il ritardo medio di arrivo dei treni e la distanza geografica tra le due stazioni?

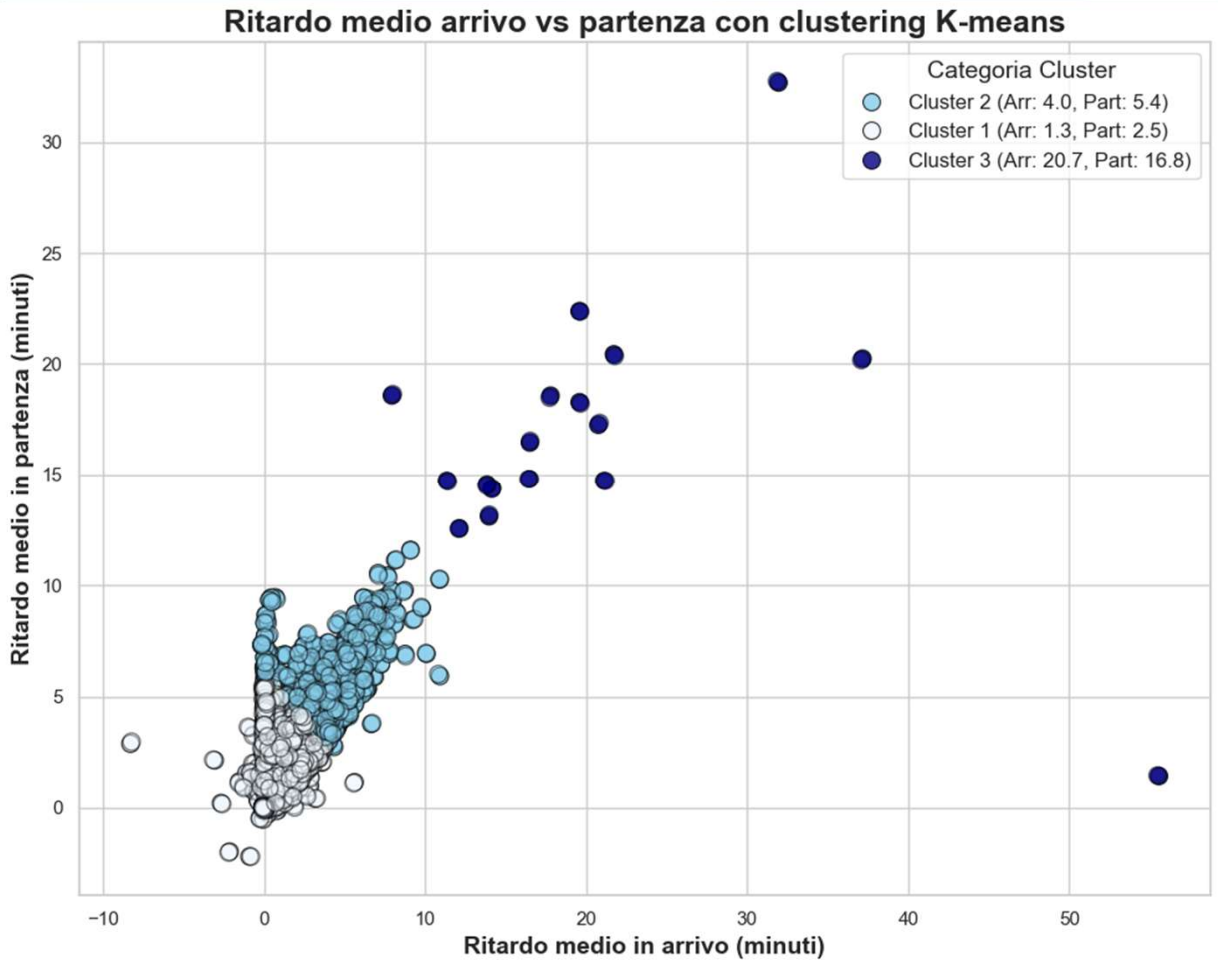
- Cosa succede se raggruppiamo le stazioni in 3 cluster?

Correlazione ritardo-distanza tratta

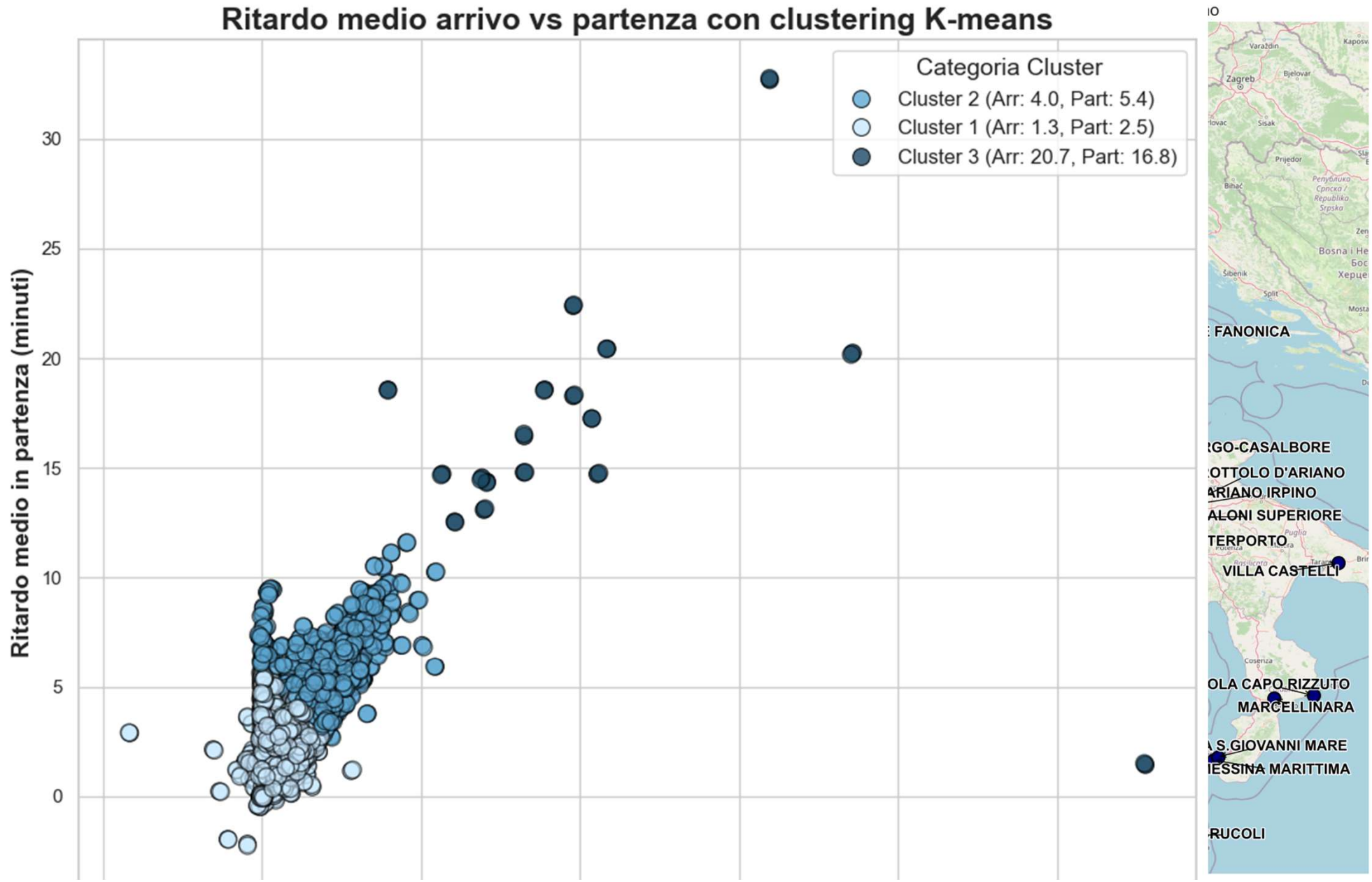


Coefficiente	Valore
Pearson	0.193
Spearman	0.191
Kendall	0.129

Ritardo medio arrivo vs partenza con clustering Kmeans (3 cluster)*



*Tratte percorse almeno 10 volte


Ritardo medio arrivo vs partenza con clustering K-means
Categoria Cluster
Cluster 2 (Arr: 4.0, Part: 5.4)
Cluster 1 (Arr: 1.3, Part: 2.5)
Cluster 3 (Arr: 20.7, Part: 16.8)
Ritardo medio in partenza (minuti)
30
25
20
15
10
5
0
FANONICA
RGO-CASALBORE
OTTOLO D'ARIANO
ARIANO IRPINO
ALONI SUPERIORE
TERPORTO
VILLA CASTELLI
OLA CAPO RIZZUTO
MARCELLINARA
S.GIOVANNI MARE
ESSINA MARITTIMA
RUCOLI